Roadrunner System Management

Gary Grider & Josip Loncaric, LANL

Dave Limpert, IBM

Oct. 18, 2007

Los Alamos NATIONAL LABORATORY EST. 1943
Operated by the Los Alamos National Security, LLC for the DOE/NNSA
LA-UR-07-7405

Will Roadrunner Phase 3 be manageable?

YES.

Roadrunner Phase 1 is manageable.

Roadrunner Phase 3 is not that different, and we have addressed those differences.

…but how and why?

Outline

- Experience with the Roadrunner Phase 1 system
- System management enhancements for the Roadrunner Phase 3 system
- Risk reduction through testing
- Hardware inventory
- Power, cooling, space
- Infrastructure, reliability, productivity
- Summary

Phase 3 system is similar to Phase 1 system

- Roadrunner Phase 1 met requirements
  - Passed over 10,000 acceptance tests
- System management software (xCAT+Warewulf) works well
  - LANL staff gained experience with xCAT+Warewulf on Roadrunner Phase 1 system
  - xCAT: IBM's HPC management solution for Linux and AIX
    - IBM is pursuing xCAT open source license, strategic move helped by partnership with LANL
    - IBM offers organization and support behind xCAT
  - Warewulf is open source (next release renamed Perceus)
- **Roadrunner Phase 3 is the same except:**
  - Each service node must boot & operate 360 Cell blades, 180 Opteron blades, and 12 I/O nodes: 4x workload
  - Compute node management: Triblade is 3 hardware components (1 Opteron blade + 2 Cell blades) managed as single logical node
  - Distribution of system services (monitoring, control, file access)
  - We've addressed these needs in the Roadrunner system management plan

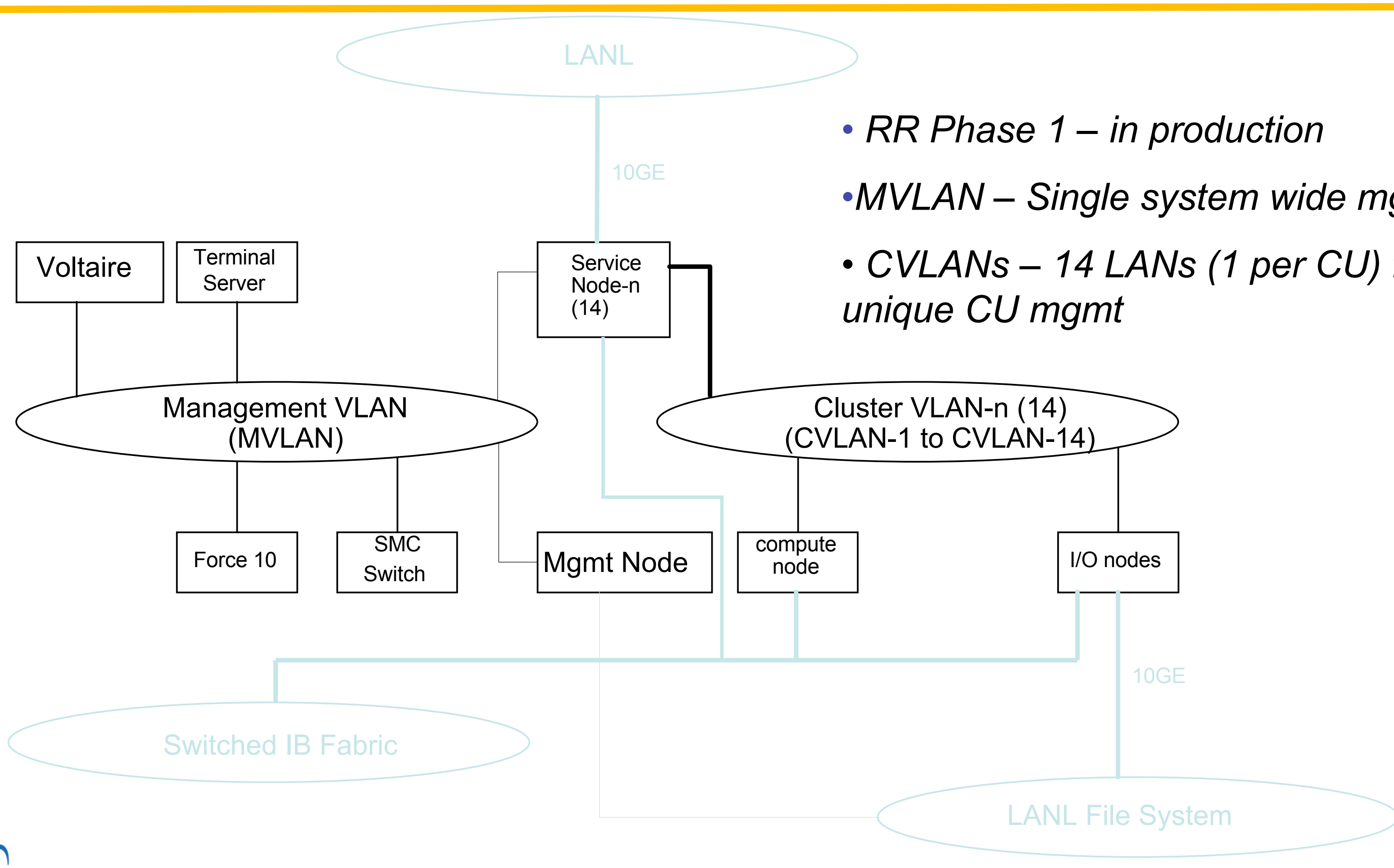
System management plan delivered per contract

- IBM / LANL partnership
- Scalable system management:
  - One Master node controls Service nodes of 18 connected units (CU)
  - Each Service node controls one CU
    - CU = service node + 12 I/O nodes + 180 triblades
      - Triblade = LS21 (Opteron) + 2*QS22 (Cell blades)
      - I/O node = x3655 (IB, 10G, 4 Opteron cores)
      - Service node = x3655 (10G, 4 Opteron cores, disk)
- Proven xCAT+warewulf clustering tools
  - Widely adopted, including in production at LANL
- Proposed enhancements tested & refined
  - Risks mitigated, details follow

Key elements of RR Phase 3 management in place

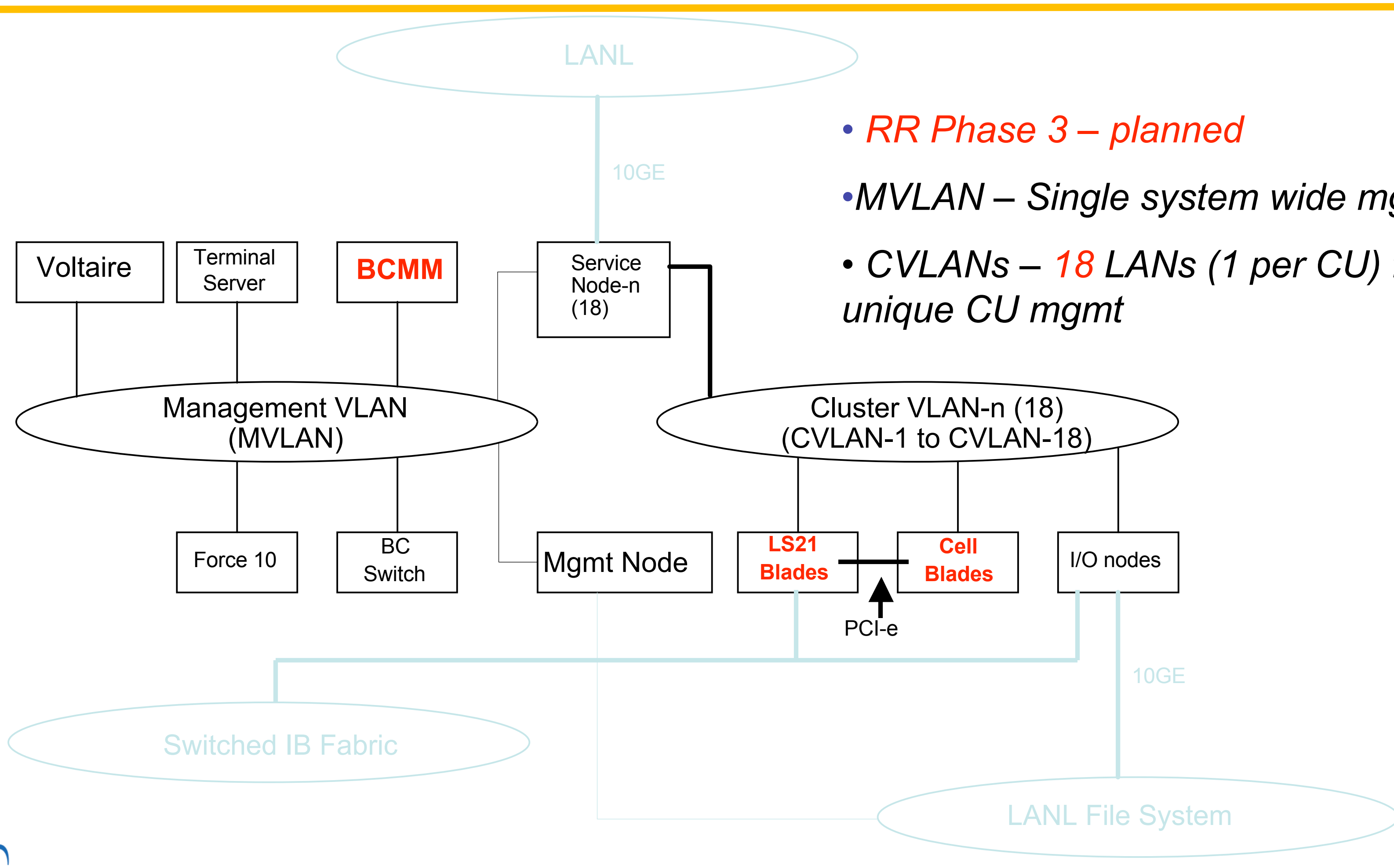
- Management of hybrid compute node’s LS21 Opteron host and QS22 Cell accelerator blades with separate physical and combined logical processes
- Management of networks necessary to support system topology
- Remote power on/off and installation/configuration, event/error detection and problem determination.

Roadrunner management networks



- *RR Phase 1 – in production*
- *MVLAN – Single system wide mgmt LAN*
- *CVLANs – 14 LANs (1 per CU) for unique CU mgmt*

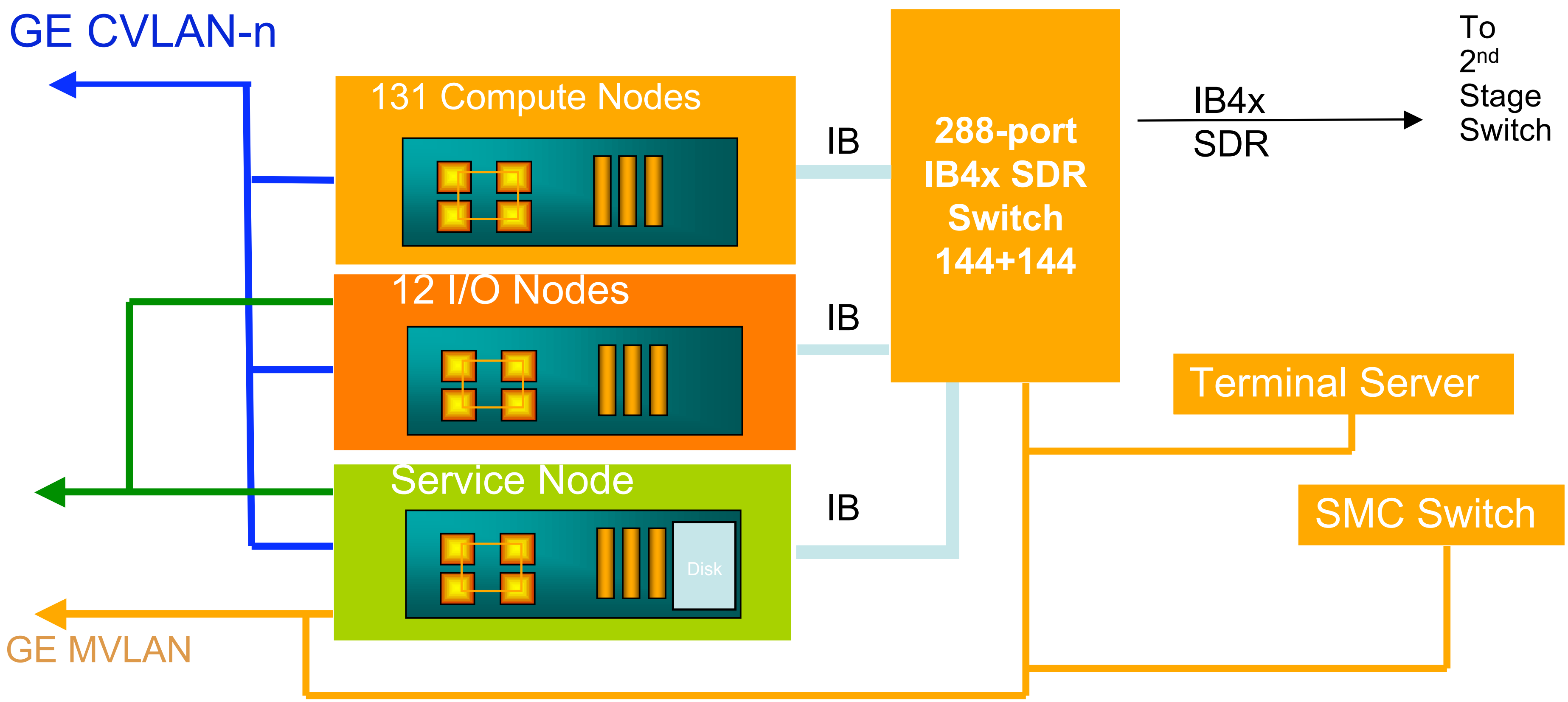
Roadrunner management networks



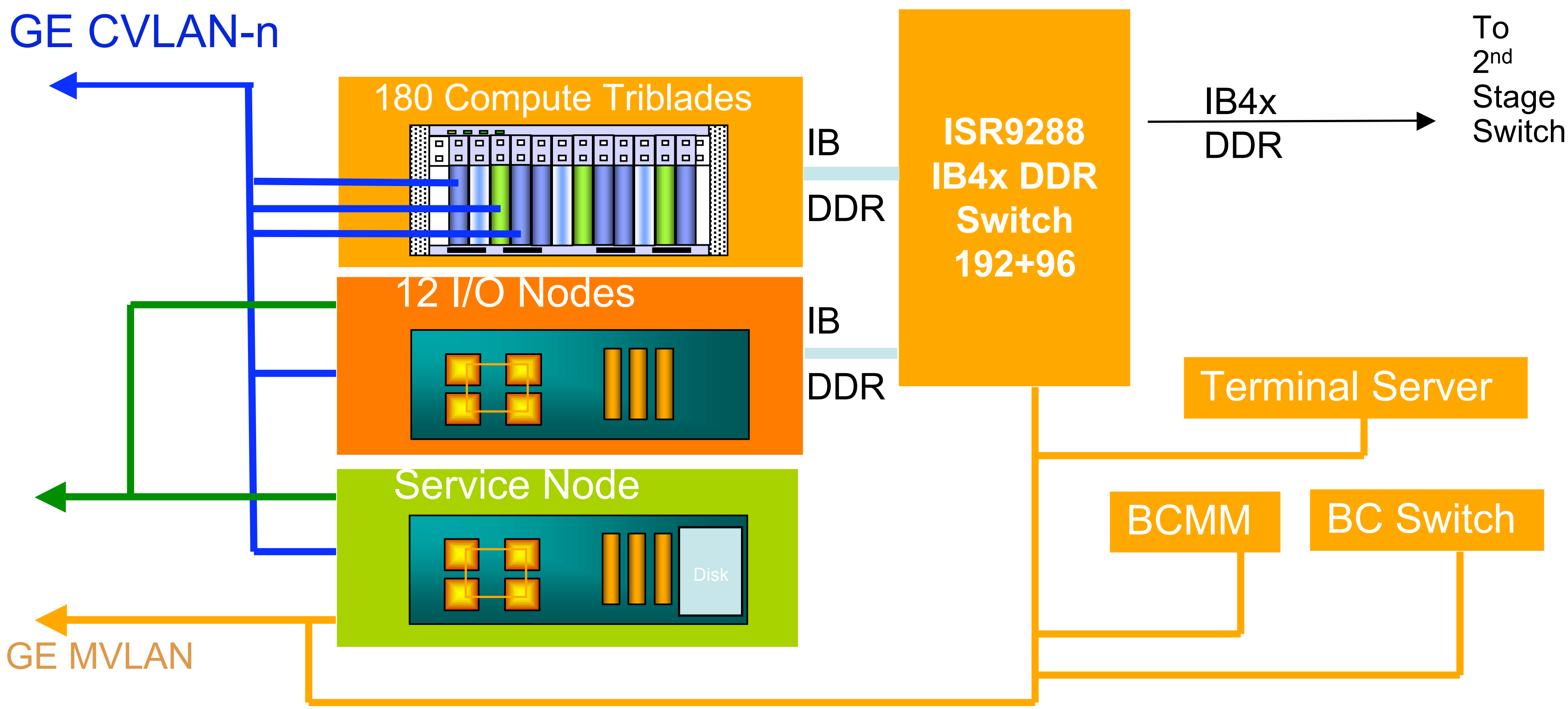
- *RR Phase 3 – planned*
- *MVLAN – Single system wide mgmt LAN*
- *CVLANs – 18 LANs (1 per CU) for unique CU mgmt*

LA-UR-07-7405

Roadrunner Phase 1 CU layout



Roadrunner Phase 3 layout is almost identical



Good I/O node performance measured, new hardware risk mitigated

- New hardware: DDR, new x3655 I/O node, new network cards
- Tested: Good BW & CPU load results on exact Phase 3 hardware:

Test	IB->10G	10G->IB	Units
Unidirectional peak	1014	985	MB/s total
Bidirectional peak	1294	1371	MB/s total
CPU load on I/O node	23%	19%	percent

- Mildly asymmetric IB->10G vs. 10G->IB, but no CPU overload
- Confirmed: I/O node performance exceeds requirements
- I/O software stack needed attention to remove CPU bottleneck

Booting process provides good performance

- Require: CU boot within 15 minutes
- Risk: 3×180+12=552 system images
  - 4x more systems per service node than Phase 1
  - Deliver possibly 100 GB over 1GbE network, >15 minutes serially
- IBM proposed mitigation: Multicast TFTP boot
  - Single physical system images sent across CVLAN network
    - One image for Cells, one image for Opterons
    - O(1) constant time method, network replicates traffic to nodes
  - Concern: Does mTFTP work reliably?
  - Trust but verify

Multicast boot operational details

- BIOS initiates PXE boot, gets small kernel using unicast
  - IBM plans to use both PXE and RFC2090 multicast mechanisms
- Client requests large RAM disk image using multicast
- Service node waits 1-30 seconds before multicast
  - Allows more clients to finish unicast stage & subscribe
  - Stragglers will pick up multicast traffic midstream
- Master client ACKs every packet until done
- Next client becomes master & requests missing packets
  - Very few packets missed, exceptional cases resolved quickly
- If no more clients, multicast TFTP is done

Multicast tested, shown as reliable as unicast, will boot large scale systems in constant time

- Test: Reliability
  - 11,500 node reboots =(100 reboots)*(115 nodes in our test cluster)
  - Failures due to PCI or memory errors, no mTFTP protocol failures
  - mTFTP protocol corrected & improved by separating unicast from multicast stages
- Test: Interference
  - NIC filters multicast
  - Nodes don't get interrupts unless they want multicast traffic
- Test: Impact on other activity
  - Node-node netperf sees impact 0.5% or less
- Key observations:
  - Separating unicast from multicast stages improves speed & reliability
    - Multicast and unicast traffic compete at switch ports
  - Booting 115 nodes takes 2-3 minutes using mTFTP, about 2-3x speedup
    - Booting Phase 3 system may take 6-9 minutes (3 distinct images)
    - Time to deliver even large 320 MB images is trivial (13.6 seconds)
    - Expect 50-100x bandwidth gain over unicast at RR Phase 3 scale
      - Parallel 2.5 GiByte/sec demonstrated, expect scaling to ~ 8+ GiByte/s
  - Boot time *doesn't* grow as number of nodes grows (constant time property)

Reliable & tight global clock synchronization

- Btime vs. NTP
  - 100x tighter clock tracking with Btime, a lightweight protocol
  - Better resistance to network delays at heavy network load
  - Btime is in production use at LANL, improves system reliability
    - Reasonably correct local clocks required by various timeout logic, timing tools
- Linux kernel 2.6.18 change
  - Timekeeping overhauled
  - New clocksource architecture, tickless kernels
  - Meaning of kernel time structures changed
  - Multi-level clock tracking
  - Btime needs to adapt
- Use NTP until Btime is ready
  - Standard tool, usable until Btime is ready

Moab scheduler & Torque resource manager

- Moab is the scheduler required by Tri-Labs
- Moab and Torque are in production use on Roadrunner Phase 1
- Function: Organize machine resources for a mix of application development and large production jobs
  - Provide control of machine, users, and jobs
  - Provide interactive and batch user entry interface to the machine
  - Provide user information on machine and job status including job accounting
  - Queue and schedule jobs according to LANL policy
  - Automatically mitigate machine failures reconfiguring resources as required.
- The service nodes will delegate actual scheduling to MOAB on an external management node.

Other concerns addressed

- Triblade RAM & NIC replacement
  - Can be done by trained LANL operators
  - Return Triblade to IBM only in case of major problems
- Triblade health monitoring
  - Separate error/alerts from Opteron and Cells in Triblade combined into single logical node status
  - Opteron-Cell status communication, Cell-Service node event stream
  - Whole Triblade reboot after failure diagnostics
  - Resource manager responsible for pre/post job cleanup

RR Phase 3 will benefit from LANL's centralized system monitoring developed for RR Phase 1

- Roadrunner Phase 1 central monitoring host collects:
  - Event tracking (syslog, snmp, alerts,…)
  - Polled temperatures/voltages/fan speeds & system load
  - Moab scheduler completed jobs data
  - Polled InfiniBand host channel adapter (HCA) status
  - Ethernet fabric anomalies
  - Asset tracking
- Monitoring of Roadrunner Phase 1 is undergoing trials
- Data filtered for presentation to multiple customers
- Objectives
  - Fast problem identification, diagnosis & correction
  - Efficient operation of more hardware without more staff
- Roadrunner Phase 3 will benefit from this technology

Roadrunner Phase 3 makes petascale possible

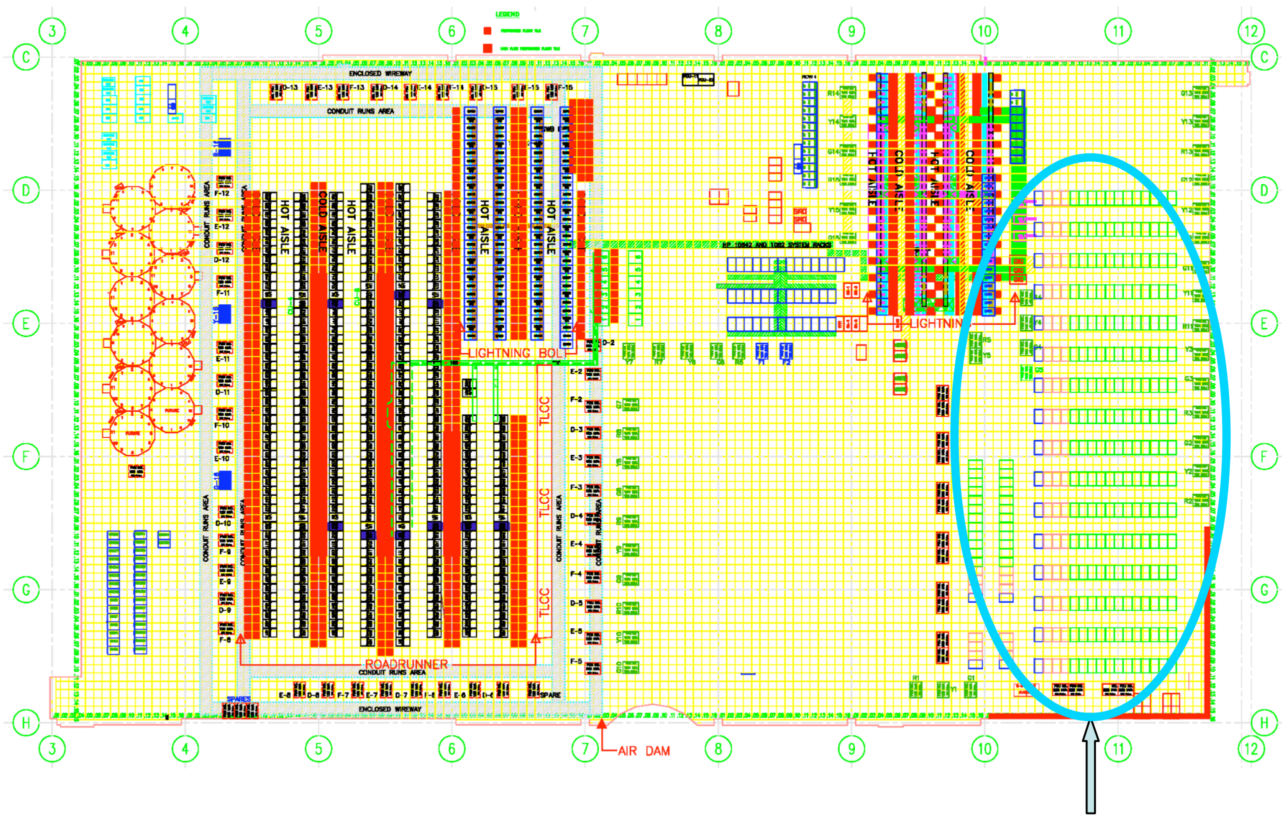
- Roadrunner is a petascale system consisting of:
  - 18 connected units, 296 racks total
    - 3240 compute triblades in 1080 chassis
    - 216 I/O nodes
    - 18 service nodes
    - 1 master node
    - 26 InfiniBand 288-port switches
    - Terminal servers, management network switches, etc.
- Roadrunner needs less than 4 MW at full load
- Roadrunner requires less than 1,135 tons of cooling
- Roadrunner footprint is 296 racks
- ***Roadrunner fits into LANL's existing facilities***
  - No further facility upgrades required

Roadrunner Phase 3 is power efficient

System	Power efficiency
Roadrunner Phase 3	0.351 TF/s per kW
BG/L	0.350 TF/s per kW
TLCC	0.197 TF/s per kW
Purple	0.016 TF/s per kW

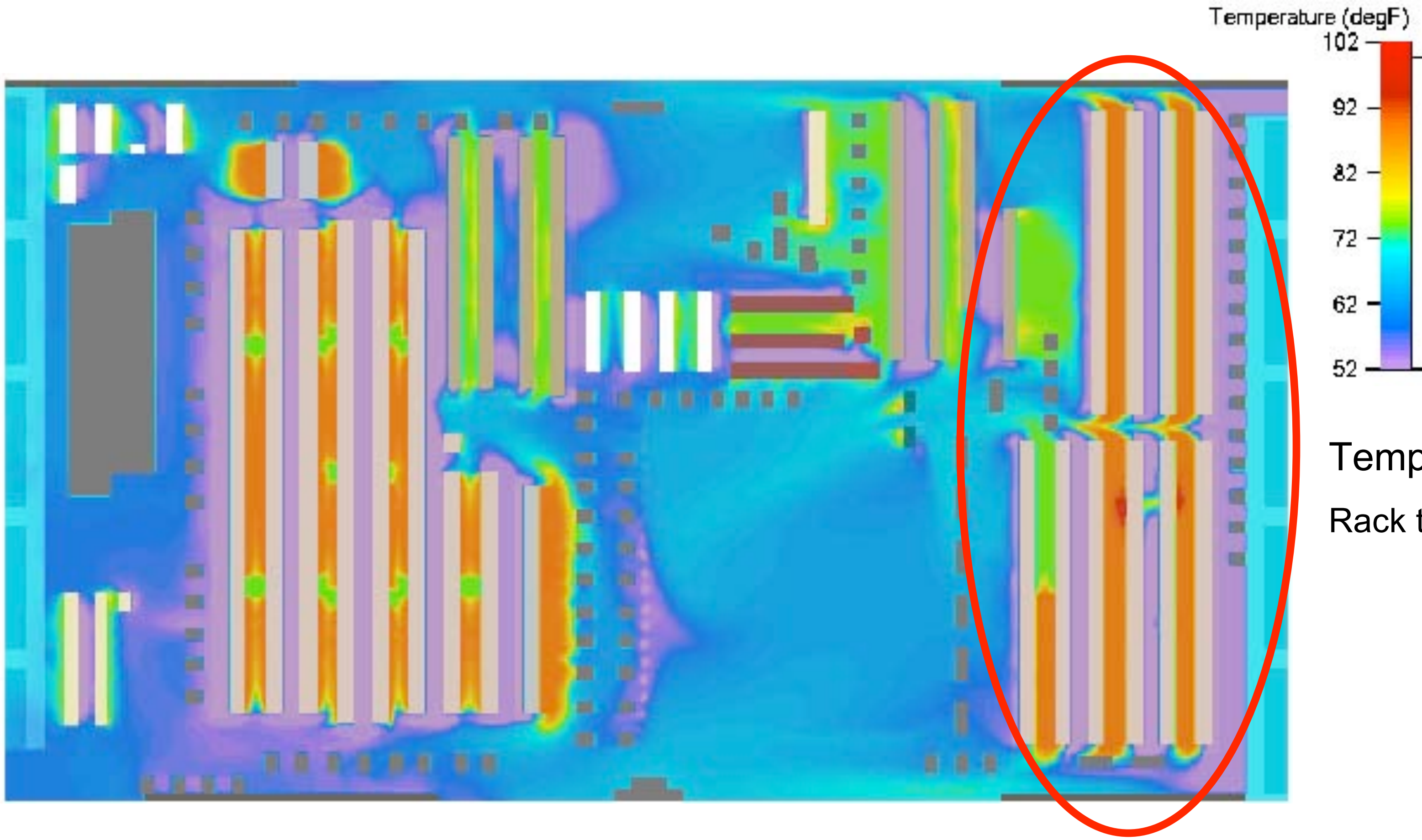
- Roadrunner allows greater performance within LANL's power budget
  - Based on IBM's very conservative power estimates
- "Petascale TLCC" system would need ~80% more power, ~50% more space, and require facility upgrades
  - TLCC = Tri-lab Linux Capacity Clusters to be delivered in 2008 (latest commodity technology)

Compact footprint in SCC



1.37 PF/s Roadrunner Phase 3

Roadrunner Phase 3 will be cooled



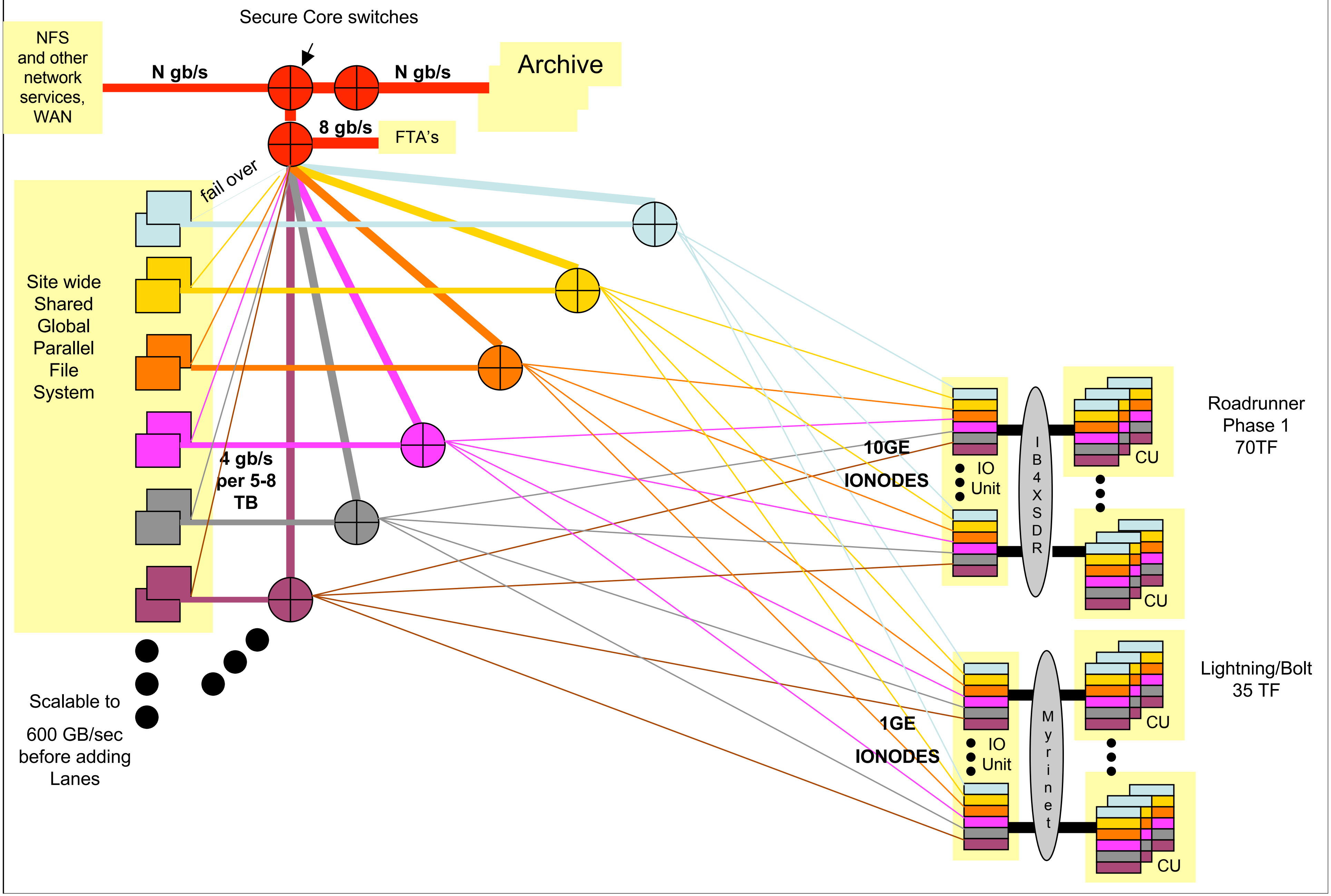
Temperature

Rack top level

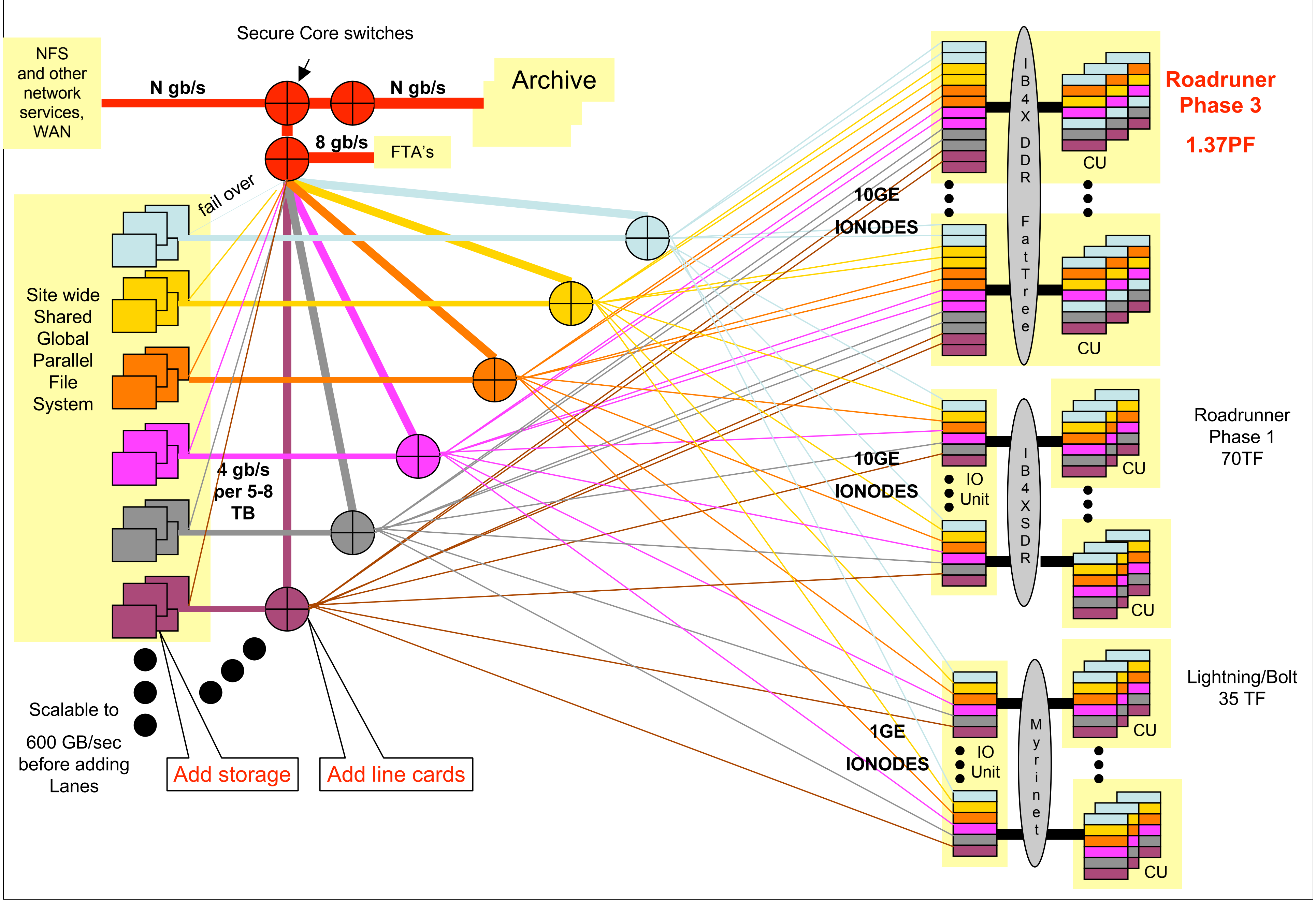
Roadrunner Phase 3 fits into LANL's scalable infrastructure design

- LANL file system and I/O SAN infrastructure is:
  - In place
  - Debugged
  - In use in production
  - Scalable
  - Low risk
- Roadrunner Phase 3 is no different than other systems:
  - Just connect to LANL infrastructure & go

Petascale Red Infrastructure Diagram (6 lanes)

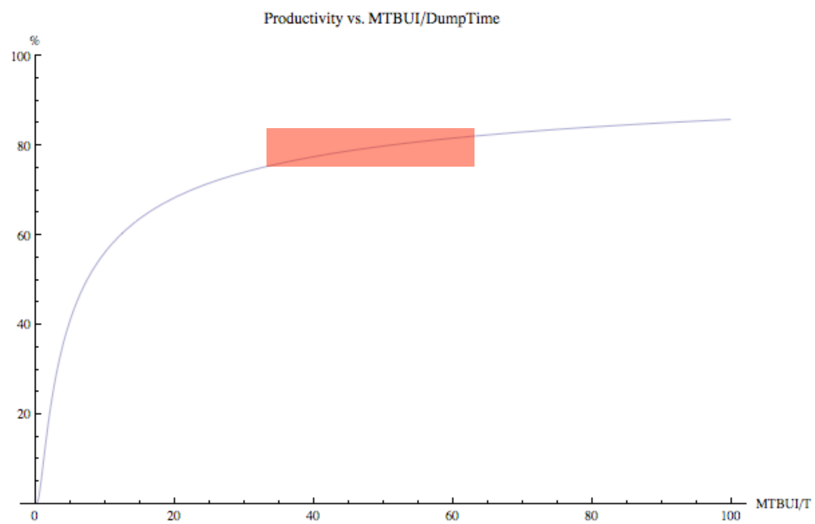


Petascale Red Infrastructure Diagram with Roadrunnner Phase 3



Roadrunner Phase 3 will deliver reasonable productivity at petascale

- Active constraint: Time to dump checkpoint files at scale
  - Productivity = SolveTime/<SolveTime+DumpTime+RestartTime+ReworkTime>
  - Don't confuse this definition of productivity with availability (availability target: 100%)
- I/O infrastructure designed to deliver good productivity at optimal checkpoint policy
  - 15-30 minutes to write RR Phase 3 full scale checkpoint files to disks per 2.5-3.5 hrs compute
  - Enough capacity for 15-30 full scale checkpoint files
  - Productivity of 75%-80% expected, based on MTBUI/DumpTime ratio
  - Target range within 5% of maximum return on investment, using optimal checkpointing policy:



This graph is universal:
Optimal productivity is determined by the ratio MTBUI/DumpTime

Under nominal assumption:

DumpTime=RestartTime

Summary: Roadrunner Phase 3 will be manageable

- It is power efficient and will be cooled efficiently
- It will fit into LANL's facilities and infrastructure
- It will deliver good I/O bandwidth, from single CU to full scale
- It will deliver reasonable productivity at petascale, cost effectively
- It will be booted quickly using multicast
- It will effectively manage Triblades as physical and logical entities
- It will be centrally monitored to identify problems quickly
- It will be serviced locally in typical failure cases

Roadrunner System Management: Abstract

- Roadrunner Phase 1 is a 70 TeraFLOP/s Opteron cluster already in production. Roadrunner Phase 3 will be a similar but larger 1.37 PetaFLOP/s cluster of hybrid Triblade nodes, each consisting of an Opteron blade and two Cell blades. System management of Roadrunner Phase 3 will address those differences, and use methods already proven on the Phase 1 system. Risks have been reduced by testing proposed system management enhancements, and showing that the Phase 3 system can be productive within LANL's power, cooling, space, and I/O infrastructure capabilities.